****

**Universidad de Chile**

Facultad de Ciencias Sociales

Profesora: Catalina Canals

Ayudantes: Daniel Cifuentes, Camila Lizama,

Daniela Olivares, Sofía Villalobos

**Pauta Control n°2**

 **Estadística II, sección 2**

**Ejercicios**

1. **A una muestra probabilística de 30 chilenos se le encuestó anualmente durante una década. Un investigador quiere utilizar dichas encuestas para analizar si las horas semanales trabajadas semanalmente disminuyeron entre el 2015 y el 2016. La información que tiene de la muestra es la siguiente (15 pts.).**

|  |  |  |
| --- | --- | --- |
|  | **Horas semanales trabajadas por los individuos en el 2015** | **Horas semanales trabajadas por los individuos en el 2016** |
| **Valor P en prueba Kolmogorov-Smirnov de normalidad** | **0,3** | **0,1** |
| **Valor P en prueba Shapiro-Wilks de normalidad** | **0,07** | **0,12** |

* 1. **Identifique la hipótesis de investigación (3 pts.)**

“las horas semanales trabajadas semanalmente disminuyeron entre el 2015 y el 2016” (o equivalente, 3 pts.).

“las horas semanales trabajadas semanalmente por la **muestra** disminuyeron entre el 2015 y el 2016” (o equivalente, 0 pts.).

* 1. **Sugiera qué test de hipótesis utilizar y por qué (6 pts.).**

Test T (1 pt.) para medias (1 pt.) de 1 cola (1 pt.) para muestras relacionadas (2 pts), referirse al tamaño muestral en la justificación (1 pt.; a diferencia del test Z sirve para n pequeño).

* 1. **Señale cuáles son todos los supuestos del test e Indique si se cumple cada uno de ellos (en caso de no hacerlo, justifique porqué a pesar de ello propone realizar dicho test) (4 pts.).**

Muestra probabilística: se cumple por enunciado (1 pt.)

Variable cuantitativa: hrs. Trabajadas es intervalar (1 pt.)

Variables normales: dado que el tamaño muestral es pequeño, corresponde utilizar el test de shapiro-wilks. Considerando los datos de la tabla, y una significación de 5%, no se rechaza la hipótesis de normalidad, por ende se cumpliría el supuesto (2 pts.).

Adicionalmente se descontó por incorporar supuestos errados.

* 1. **Plantee las hipótesis nula y alternativa del test a realizar (2 pts.).**

H1: “las horas semanales trabajadas semanalmente disminuyeron entre el 2015 y el 2016” (o equivalente, 1 pts.).

H0: “las horas semanales trabajadas semanalmente se mantuvieron constantes entre el 2015 y el 2016” (o equivalente, 1 pts.).

Si las hipótesis hacían referencia a la muestra: 0 pts.

1. **Un grupo de investigadores estudia la importancia de lo geográfico en las decisiones educativas de las familias. Específicamente, a partir de una muestra probabilística de 5000 estudiantes del Gran Santiago, compara el promedio de distancia que los estudiantes de NSE Bajo recorren para ir al colegio, y el promedio de distancia que los estudiantes de NSE Alto recorren para ir al colegio. Su hipótesis de investigación es que los estudiantes de mayor nivel socioeconómico asisten a colegios más distantes de sus hogares. Considerando dicha hipótesis como hipótesis alternativa, y considerando como hipótesis nula que las distancias estudiante-escuela son iguales en ambos niveles socioeconómicos, los investigadores realizan un Test Z para diferencia de medias, estimando el siguiente estadístico:**

$$Z\_{obs}=\frac{\overbar{X}\_{B}-\overbar{X}\_{A}}{SE(\overbar{X}\_{B}-\overbar{X}\_{A})}=-2,5$$

**donde** $\overbar{X}\_{B}$ **es la distancia estudiante-escuela para los estudiantes de nivel socioeconómico bajo, y** $\overbar{X}\_{A}$ **para los de nivel socioeconómico alto (14 pts.).**

* + - * 1. **Estime e interprete el Valor P. Muestre gráficamente a qué probabilidad corresponde el Valor P (explique brevemente su gráfico) (7 pts.).**

El Valor P=P[Z<-2.5]=0,6% (2 pts.). Esto quiere decir que la probabilidad de, si H0 es cierta, obtener un Z=-2,5 o uno más contradictorio con H0 (es decir, un Z más pequeño) que éste, corresponde a 0,6% (3 pts.).

El siguiente gráfico muestra el valor P: la curva muestra la distribución normal estándar del estadístico en el caso que la H0 sea cierta, y la zona achurada (valor P), muestra cuál era la probabilidad de que siendo H0 cierta obtuviésemos el valor que obtuvimos o uno más pequeño (2 pts.).



* + - * 1. **Con 5% de significancia, ¿Rechazaría su hipótesis nula? Interprete sus resultados sociológicamente (7 pts.).**

Dado que el valor p es menor que la significancia se rechaza H0 (3 pts.)

Explicar con claridad al menos una hipótesis pertinente (4 pts.).

1. **Posteriormente los investigadores se centran en estudiar las comunas de Santiago, El Bosque y Recoleta. Particularmente analizan si la distancia recorrida por los estudiantes para llegar al colegio varía según comuna. Para ello analizan un ANOVA de una vía, obteniendo los siguientes resultados.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Suma de Cuadrados** | **Grados de Libertad** | **Medias de cuadrados** | **F** | **Prob>F** |
| **Total** | **54.000** | **1000** | **54** |  |  |
| **Comuna** | **4.000** | **2** | **2.000** | **40** | **0,0002** |
| **Residual** | **50.000** | **1000** | **50** |  |  |

|  |  |
| --- | --- |
|  | **Eta cuadrado** |
| **Comuna** | **7,4%** |

**A partir de la tabla los investigadores sacan las siguientes conclusiones. Identifique sus errores, explique por qué son errores y cómo los corregiría.**

**“Los resultados de la prueba ANOVA de una vía nos permite concluir, con 5% de significancia, que existen diferencias entre las varianzas de los tres grupos, en tanto el Valor P<0,05. Además, el Eta cuadrado nos permite analizar la magnitud del efecto de la variable comuna en la variable distancia estudiante-escuela; al ser este un 7,4%>5%, con 95% de confianza, afirmamos que las distancias estudiante-escuela son distintas en las tres comunas. Finalmente, para analizar si el promedio de distancia recorrida es distinto entre Recoleta y Santiago, Santiago y El Bosque, y El Bosque y Recoleta, recomendamos utilizar 3 test T para muestras independientes (uno para cada par de comunas), lo cual permitirá mantener constante el error de tipo I a la hora de evaluar la hipótesis nula de que las tres comunas tienen el mismo promedio” (12 pts.).**

Error 1: “Los resultados de la prueba ANOVA de una vía nos permite concluir, con 5% de significancia, que existen diferencias entre las varianzas de los tres grupos” esto es un error porque ANOVA no es una prueba para comparar varianzas (2 pts.). Lo correcto sería decir: “Los resultados de la prueba ANOVA de una vía nos permite concluir, con 5% de significancia, que las medias de al menos dos comunas son distintas” (2 pts.).

Error 2: “el Eta cuadrado nos permite analizar la magnitud del efecto de la variable comuna en la variable distancia estudiante-escuela; al ser este un 7,4%>5%, con 95% de confianza, afirmamos que las distancias estudiante-escuela son distintas en las tres comunas”, esto es un error porque el eta cuadrado es un estimador del tamaño del efecto de una variable en otra, y no un test de hipótesis, por ende no tiene sentido compararlo con la significancia (2 pts.). Lo correcto sería decir: “el Eta cuadrado nos permite analizar la magnitud del efecto de la variable comuna en la variable distancia estudiante-escuela; al ser este un 7,4% esto quiere decir que la variable comuna explica el 7,4% de la varianza de la variable distancia estudiante-escuela” (2 pts.).

Error 3: “para analizar si el promedio de distancia recorrida es distinto entre Recoleta y Santiago, Santiago y El Bosque, y El Bosque y Recoleta, recomendamos utilizar 3 test T para muestras independientes (uno para cada par de comunas), lo cual permitirá mantener constante el error de tipo I a la hora de evaluar la hipótesis nula de que las tres comunas tienen el mismo promedio”. Esto es un error dado que el test T no mantiene constante el error tipo I al hacer múltiples comparaciones (2 pts.), si se quisieran comparar los pares de comunas sería pertinente un test de Tukey (2 pts.), dado que no tiene este problema.

1. **Un grupo de investigadores estiman chi-cuadrado de asociación para poner a prueba la hipótesis nula de que no existe asociación entre las variables dummy consume alcohol (0: No, 1: Si) y tener hijos (0: No, 1: Si) en hombres de menos de 40 años. Su estadístico observado corresponde a** $χ^{2}=5$**. La siguiente tabla muestra la distribución de la muestra en ambas variables (19 pts.).**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tiene hijos | No tiene hijos | Total |
| Consume Alcohol | 15 | 30 | 50 |
| No Consume Alcohol | 35 | 20 | 50 |
| Total | 50 | 50 | 100 |

* + - * 1. **¿Cuáles son los supuestos de la prueba Chi-cuadrado? ¿Se cumplían en este caso? Justifique su respuesta (5 pts.).**

Muestra probabilística: se cumple por enunciado (1 pt.)

Variables cualitativas: se cumple dado que ambas son dummy (1 pts.)

Menos 20% de fe<5: se cumple dado que todas las fe=25 (3 pts.).

Se descontó por incluir supuestos errados.

* + - * 1. **Indique cómo distribuye el estadístico Chi-cuadrado en este caso. Estime e interprete el Valor P (6 pts.).**
* El estadístico chi cuadrado, si H0 es cierta distribuye $χ\_{1 gl.}^{2}$ (2 pts).
* Valor P=P($χ\_{1 gl.}^{2}$>5]=2,5% (2 pts.)
* Esto quiere decir que la probabilidad de, si H0 es cierta, obtener un estadístico=5, o uno más contradictorio con H0 (esto es, más grande aún), es igual a 2,5%.
	+ - * 1. **Con 5% de significancia, ¿Rechazaría su hipótesis nula? Interprete sus resultados estadísticamente (3 pts.).**

Si, dado que Valor P<significancia. Esto implica que existe asociación entre ambas variables.

* + - * 1. **Estime V de Cramer e interprete sus resultados estadística y sociológicamente (5 pts.).**

V=$\sqrt{\frac{χ\_{ obs}^{2}}{n(k-1)}}=\sqrt{\frac{5}{100(2-1)}}=0,22$ (1 pt.)

La asociación es de baja magnitud (2 pts.).

Explicación clara de una hipótesis pertinente sobre porqué existe asociación y/o porqué esta es baja (2 pts.)

* + - * 1. **PREGUNTA OPTATIVA: ¿Cree que sería pertinente controlar por una tercera variable? ¿Por cuál? Justifique su respuesta (5 pts.).**
* Si, por una variable que creen que incide en el consumo de alcohol (1-3 pts. Según justificación).
* Si, por una variable que creen que incide en el consumo de alcohol, y que es tal que, podría asociarse con “no tener hijos”, pero no con “tener hijos”, de modo que en realidad esta variable se relacione con el consumo de alcohol y no tener hijos (5 pts). Ejemplo:
	+ Controlar por edad (suponiendo que el estudio solo considera a mayores de edad, se sugieren tramos: 18-25 años, 26-33 y 34-40). Si usamos esta variable, podría suceder que en realidad no es que tener hijos se asocie a un menor consumo de alcohol, sino que en realidad, la edad se asocia al consumo de alcohol (consumiendo menos, al pertenecer a un tramo de mayor edad), de modo que en los hombres sin hijos, son más los que consumen alcohol no porque no tengan hijos, sino porque en general los hombres sin hijos son más jóvenes, mientras que, en los hombres con hijos, son menos los que consumen alcohol, no porque tengan hijos, sino porque en general los hombres con hijos tienen más edad, y mayor edad se asocia a menos consumo de alcohol. Esta hipótesis sería corroborada, si al hacer el análisis para cada tramo etario por separado la asociación fuese nula.